

Ngôn ngữ học ngữ liệu

Parsing

Quy Nguyen

Nội dung

1. Constituency
2. Dependency

Constituency syntax

- ▶ Nhãn ngữ (constituency labels): cho biết loại cú pháp của ngữ
- ▶ Nhãn chức năng (functional labels): cho biết chức năng và ngữ nghĩa (nếu có) của ngữ
- ▶ Nhãn phần tử rỗng (null elements): dùng để làm dấu các thành phần khuyết trong câu
- ▶ Chỉ số tham chiếu (reference indices): để lưu vết di chuyển của các thành phần cú pháp

Nhãn ngữ

No.	Tag	Meaning
1	NP	Noun phrase
2	QP	Quantitative phrase
3	VP	Verb phrase
4	ADJP	Adjective phrase
5	PP	Prepositional phrase
6	RP	Adjunct phrase
7	CONJC	Conjunction phrase
8	UCP	Unlike coordinated phrase
9	QNP	Questioning noun phrase
10	QADJP	Questioning adjective phrase
11	QRP	Questioning adjunct phrase
12	QPP	Questioning prepositional phrase
13	MDP	Modal phrase

Nhãn câu/mệnh đề

No.	Tag	Annotation
1	S	Simple sentence, compound sentence, or complex sentence
2	SQ	Question
3	SPL	Special sentence
4	SBAR	Relative clause

Nhãn chức năng

No.	Tag	Annotation
1	H	Head of phrase
2	SBJ	Subject
3	LGS	Logical subject
4	PRD	Predicate that is not VP
5	DOB	Direct object
6	IOB	Indirect object
7	CMP	Complement
8	TPC	Topicalized
9	MDP	Modal phrase
10	VOC	Vocative
11	PRN	Parenthetical
12	TMP	Temporal
13	LOC	Locative
14	MNR	Manner
15	PRP	Purpose or reason
16	CND	Condition
17	CNC	Concessions
18	ADV	Adverbial
19	HLN	Headline
20	TTL	Title
21	EXC	Exclamative sentence
22	CMD	Imperative sentence

Nhãn phần tử rỗng

No.	Tag	Nghĩa
1	*T*	Trace of movement
2	*E*	Ellipses without trace
3	*	Ellipses with trace
4	*0*	Null complementizer
5	*P*	Null passive verb
6	*H*	Ellipses with trace for head word
7	*D*	Ellipses with trace for direct object of reduced relative clause
8	*SUM*	Null post-modifier which is one of components of an aggregate

Nhãn câu/mệnh đề

- ▶ S: gán cho câu đơn, câu ghép
- ▶ Ví dụ:

Tôi đi học.

(S Tôi đi học
(PU .))

Mây tan rồi mưa tạnh.

(S (S Mây tan)
(Cp rồi)
(S mưa tạnh)
(PU .))

Nhãn câu/mệnh đề

- ▶ SQ: gán cho câu hỏi

Ví dụ: **Anh hay tôi làm?**

(SQ (Nn Anh) (Cp hay) (Pp tôi) (Vv làm)
(PU ?))

- ▶ SPL: gán cho câu có cú pháp đặc biệt

Ví dụ: Câu tồn tại

Có 3 người trong phòng.

(SPL (Ve-H Có) 3 người trong phòng)
(PU .))

Nhãn câu/mệnh đề

► SBAR: Mệnh đề phụ kết

Ví dụ:

Nó đã đọc (NP-DOB *quyển sách* (SBAR (Cs mà) (S anh đưa))).

Nó đã đọc (NP-DOB *quyển sách* (SBAR (S anh đưa))).

Nhãn ngữ

► Ngữ danh từ gán NP

Ví dụ:

(NP-SBJ Nó) đã đọc (NP-DOB quyển sách (SBAR (Cs mà) (S anh đưa))).

► Cụm số từ QP

Ví dụ:

tôi có (NP 2/Num con mèo) → không gán QP cho số "2" vì chỉ có 1 từ

tôi được (NP (QP trên 5) điểm toán)

Anh ấy ăn (NP (QP những/M sáu) bát cơm) Nó được bao nhiêu điểm? Chắc (QP bảy/Num-H hơn/R)

Nhãn ngữ

- ▶ Ngữ động từ gán VP

Ví dụ:

Nó (VP đang đi rất nhanh về phía cô ta).

Nó (VP cười/Vv-H hi_hì/ON)

- ▶ Ngữ tính từ gán ADJP

Ví dụ:

Cuộc đàm phán (ADJP (RP sẽ/R rất/R) khó_khăn/Aa)

Nó (ADJP giỏi/Aa-H toán).

Nhãn ngữ

► Ngữ giới từ gán PP

Ví dụ:

Dĩ trái cây này (PP (ADJP không/Ra phải/Aa) để/Cs-H (VP ăn)).

Tôi đi học (PP bằng (NP xe bus))

► Ngữ phụ từ gán RP

Ví dụ:

Chúng ta (VP (RP sẽ/R không/R-H) đi chơi vào tuần tới).
Ai (ADJP (RP rồi/R cũng/R-H sẽ/R) già đi).

Nhãn ngữ

- ▶ Cụm kết từ gán **CONJP**

Ví dụ:

(*CONJP (Cp-H cùng) (Cp-H với)*)

(*CONJP (Cp-H và) (Cp-H cùng) (Cp-H với)*)

- ▶ Nhiều ngữ khác loại nhưng có cùng chức năng cú pháp
(Unlike coordinated phrase) gán **UCP**

Ví dụ:

Cô ấy đẹp và nấu ăn ngon.

(*S (NP-SBJ Cô ấy)*

(*UCP-PRD (ADJP (Aa-H đẹp)*

(*Cp và)*

(*VP (Vv-H nấu_ăn) (Aa ngon)))*

(*PU .)*)

Nhãn ngữ

- ▶ Cụm từ để hỏi: **Thêm "Q" vào trước cụm danh từ, cụm giới từ** có bao gồm đại từ để hỏi như *ai, gì, sao, nào, mấy, thế_nào, bao_nhiều, bao_giờ, bao_lu, Ở_u, ...*

Ví dụ:

Hôm nay chúng ta ăn (QNP gì/Pp-H)?
cuốn sách này (QPP của/Cs-H (NP ai))?
(QNP Khi/Nt (NP nào)) bạn về quê?

- ▶ Cụm từ cảm thán (Modal phrase) gán **MDP**

Ví dụ:

(MDP Trời ạ), tôi làm sao biết được!
(S (MDP A) (PU !)) (NP-SBJ mẹ) (VP đã về))

Nhãn ngữ

- ▶ Các cụm từ chưa được định nghĩa loại thì gán **XP**

Ví dụ:

(S (**XP âm âm âm**) (PU ,) (NP-SBJ *những tiếng nổ mìn*) (VP *vang vọng*) (PU .))

Nhãn ngữ - Ví dụ

- ▶ Đất nước đã bị máy bay Mỹ tàn phá.
- ▶ SCIC tiếp tục thoái vốn toàn bộ tại 2 doanh nghiệp lớn.
- ▶ Hai khoản đầu tư trên có giá trị tương ứng là 180 tỉ và 176 tỉ đồng.
- ▶ Những dự án đã và đang bồi thường, hỗ trợ thiệt hại và tái định cư vẫn tiếp tục thực hiện theo phương án bồi thường cũ.

Constituency syntax - Example

- ▶ Câu tiếng Việt:
Đất nước ta đã bị máy bay Mỹ tàn phá.
- ▶ Kết quả WS và POS tagging:
*Đất_nước/Nn ta/Pp đã/R bị/Vv máy_bay/Nn Mỹ/Nr
tàn_phá/Vv ./PU*
- ▶ Gán nhãn câu **S**, chủ ngữ **NP**, vị ngữ **VP**
(**S** (**NP** (Nn Đất_nước) (Pp ta))
 (**VP** (R đã) (Vv bị) (Nn máy_bay) (Nr Mỹ) (Vv tàn_phá))
 (PU .))

Constituency syntax - Example

- ▶ Gán nhãn câu **S**, chủ ngữ **NP**, vị ngữ **VP**

(S (NP (Nn Đất_nước) (Pp ta))
 (VP (R đã) (Vv bị) (Nn máy_bay) (Nr Mỹ) (Vv tàn_phá))
 (PU .))

- ▶ Tiếp tục phân tích chủ ngữ, vị ngữ, gán nhãn từ trung tâm của ngữ, head word H

(S (NP (Nn-H Đất_nước)
 (NP(Pp-H ta)))
 (VP (R đã) (Vv-H bị)
 (S (NP (Nn-H máy_bay)
 (NP (Nr-H Mỹ)))
 (VP (Vv-H tàn_phá))))
 (PU .))

Constituency syntax - Example

- ▶ Gán nhãn phần tử rỗng - Lưu vết sự dịch chuyển: ***T***

(S (NP-SBJ-1 (Nn-H Đất_nước)
 (NP(Pp-H ta)))
 (VP (R đã) (Vv-H bị)
 (S-CMP (NP-SBJ (Nn-H máy_bay)
 (NP (Nr-H Mỹ)))
 (VP (Vv-H tàn_phá)
 (NP-DOB *T*-1)))
 (PU .))

Constituency syntax - Example

- ▶ **SCIC tiếp tục thoái vốn toàn bộ tại 2 doanh nghiệp lớn.**

(S-HLN
 (NP-SBJ (Nr-H SCIC))
 (VP (Vv-H tiếp_tục)
 (VP-CMP (Vv-H thoái)
 (NP-DOB (Nn-H vốn) (Nw toàn_bộ)
 (PP (Cs-H tại)
 (NP (Num 2) (Nn-H doanh_nghiệp)
 (ADJP (Aa-H lớn))))))))
(PU .))

Constituency syntax - Example

- ▶ Hai khoản đầu tư trên có giá trị tương ứng là 180 tỉ và 176 tỉ đồng.

(S
 (NP-SBJ (Num hai) (Nn-H khoản) (Vv đầu_tư) (Nn
trên))
 (VP (Ve-H có)
 (S-CMP
 (NP-SBJ (Nn-H giá_trị) (Vv tương_ứng))
 (VP (Vc-H là)
 (NP-CMP
 (QP
 (QP (Num-H 180) (Nq-H tỉ))
 (Cp và)
 (QP (Num-H 176) (Nq-H tỉ)))
 (Nu-H đồng))))))
 (PU .))

Constituency syntax - Example

(S
 (NP-SBJ-1 (Nq Những) (Nn-H dự_án)
 (VP
 (VP
 (RP (R-H đã) (Cp và) (R-H đang))
 (Vv-H bồi_ thường)
 (PU ,)
 (Vv-H hỗ_ trợ)
 (NP (Nn-H thiệt_ hại)))
 (Cp và)
 (VP (Vv_ w (Sv tái) (Vv-H định_ cư))))))
 (VP (R vẫn) (Vv-H tiếp_ tục)
 (VP-CMP (Vv *P*)
 (VP-CMP (Vv-H thực_ hiện)
 (NP-DOB (NONE *T*-1))
 (VP-MNR (Vv-H theo)
 (NP (Nn-H phương_ án) (Vv bồi_ thường) (Aa cũ))))))
 (PU .))

Nhãn chức năng

- ▶ SBJ: được gán cho thành phần chủ ngữ (**subject**) của câu/mệnh đề
- ▶ DOB, IOB và CMP: nhãn chức năng của object của động từ
 - ▶ Ngoại động từ (transitive verb): Chủ thể (subject) hành động (transitive verb) và gây ra hoặc tác động đến chủ thể khác (object).
Anh ấy_{subject} pha_{transitiveV} cà phê_{directO} cho chúng tôi_{indirectO}
 - ▶ Nội động từ (Intransitive verb): hành động nội tại của người nói hay người, chủ thể thực hiện hành động (Subject). Hành động của Subject đó không tác động trực tiếp lên bất kỳ Đối tượng (Object) nào. Nói cách khác Intransitive Verb không có hoặc không cần Direct Object (Bổ ngữ trực tiếp) để tạo thành 1 câu có nghĩa. Intransitive verb không thể chuyển sang thể bị động được. Intransitive verb thường là những Động từ chỉ hành động như: arrive, go, lie, sneeze, sit, die, run, faint, hesitate, occur, pause, rain, remain, sleep, etc.

Nhãn chức năng

- ▶ SBJ: được gán cho thành phần chủ ngữ (**subject**) của câu/mệnh đề
- ▶ Object (DOB, IOB): được gán cho object của transtive verbs
 - ▶ DOB: Chủ thể (subject) tác động (ngoại động từ - transtive verb) trực tiếp đến chủ thể khác (Direct object).
 - ▶ IOB: Người/vật được thụ hưởng

(S (NP-SBJ Tôi)
 (VP tặng
 (NP-IOB cô ấy)
 (NP-DOB hoa))
(PU .))

(S (NP-SBJ Tôi)
 (VP tặng
 (NP-DOB hoa)
 (PP-IOB cho
 (NP cô ấy))))
(PU .))

Nhãn chức năng

- ▶ CMP (Complement): Gán cho loại object còn lại (không phải DOB hay IOB)

(S (NP-SBJ Tôi)
 (VP là/Vc-H
 (NP-CMP sinh viên))
(PU .))

(S (NP-SBJ Tôi)
 (VP có
 (NP-CMP rất nhiều tiền))
(PU .))

Constituency syntax - Example

- ▶ Gán nhãn chức năng cú pháp của ngữ
 - ▶ SBJ: được gán cho thành phần **subject** của câu/mệnh đề

(S (NP-SBJ (Nn-H Đất_nước)
 (NP(Pp-H ta)))
 (VP (R đã) (Vv-H bị)
 (S-CMP (NP (Nn-H máy_bay)
 (NP (Nr-H Mỹ)))
 (VP (Vv-H tàn_phá))))
(PU .))

Dependency syntax

- ▶ Universal Dependencies 2.6:
<https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3226>
- ▶ Vietnamese UD treebank is a automatic conversion of the constituent treebank created in the VLSP project automatically converted to UD

POS Tags

- ▶ Universal Dependency bao gồm 17 nhãn
- ▶ Tuy nhiên, UD Vietnamese chỉ sử dụng 14 trong số 17 nhãn này.

Không sử dụng:

- ▶ ADV (adverb): Adverbs are words that typically modify verbs for such categories as time, place, direction or manner.

Ex: very, well, exactly, ...

- ▶ SYM (symbol):

\$, %, §, ©, +, , , , =, <, >, :)

- ▶ PRON (pronoun):

personal pronouns: I, you, he, she, it, we, they

reflexive pronouns: myself, yourself, himself, herself, itself, ourselves, yourselves, themselves

interrogative pronouns: who, what as in *What* do you think?

relative pronouns: who, what as in I wonder *what* you think.

indefinite pronouns: somebody, something, anybody, anything

total pronouns: everybody, everything

negative pronouns: nobody, nothing

possessive pronouns: mine, yours, (his), hers, (its), ours, theirs

POS Tags

► **ADJ:** tính từ

Ví dụ:

đơn giản nhất là bám theo cả chục kilômet , khi nào khách mua *mới* thôi .

Hai kịch bản *mới* phải thực hiện.

Chúng em thường gọi đùa Hiên là nàng tiên *nhỏ*.

► **ADP:** Giới từ

Ví dụ:

Bác Hồ ủng hộ quyết định *của* tôi.

Ông đã trở thành người bạn thân thiết *với* toàn thể nhân dân VN .

chủ quán hện như đình đóng cột *với* Bình rằng mai , kia quay lại đây sẽ có .

Nó không *với* tới.

POS Tags

► **AUX:** Auxiliary

Ví dụ:

► English:

Tense auxiliaries: *has* (done), *is* (doing), *will* (do)

Passive auxiliaries: was (done), got (done)

Modal auxiliaries: *should* (do), *must* (do)

Verbal copulas: He *is* a teacher.

► Vietnamese:

Đó *là* cuộc chiến tranh phi nghĩa.

Hùng có thể *là* một trường hợp.

► **CCONJ:** coordinating conjunction

Ví dụ:

Người mua thì không quan tâm hàng nhập lậu *hay* nội địa

Nhưng miệng *thì* rất hẹp , đường kính chừng 6 tấc .

Thì hiện tại Chọn đúng lúc đúng *thì*

POS Tags

- ▶ DET: determiner

Ví dụ:

- ▶ English:

articles: *a, an, the*

possessive determiners: *my, your*

demonstrative determiners: this as in I saw *this* car yesterday.

interrogative determiners: which as in “*Which* car do you like?”

relative determiners: which as in “I wonder *which* car you like.”

quantity determiners (quantifiers): indefinite *any*, universal: *all*, and negative *no* as in “We have no cars available.”

- ▶ Vietnamese quantifiers:

hải và chị Hằng tìm **mọi** cơ hội để xây dựng tình cảm cho hai thanh niên.

Nhưng Lan phải thành thật kể hết **mọi** chuyện cho anh nghe. những mặt hàng điện tử cũ đã có qui định cấm nhập khẩu dưới **mọi** hình thức.

POS Tags

- ▶ INTJ (interjection: thán từ): A word that is used most often as an exclamation or part of an exclamation. It typically expresses an emotional reaction.

Ví dụ:

Hùng giật mình : “ Sao tôi không biết ông **nhỉ** ? ” .

Ai đây **nhỉ** ? ” .

POS Tags

- ▶ **PART:** Particles are function words that must be associated with another word or phrase to impart meaning.

Ví dụ:

- ▶ English:

Possessive marker: 's

Negation particle: not;

Question particle: Adding this particle to the end of a clause turns the clause into a question;

Sentence modality: *Let's* do it! *If only* I could do it over.

- ▶ Vietnamese:

Ai đấy **nhỉ** ? “ .

Thay vào đó là những bức ảnh nói về **chính** những đứa trẻ lang thang .

POS Tags

- ▶ NOUN: noun

Ví dụ:

Hai **ngày** nữa mẹ Tùng sẽ mua xe .

- ▶ NUM: numeral

Ví dụ:

Đòi lại **một** cuộc đời

- ▶ PROPN: proper noun

Ví dụ:

ý tưởng về một đàn xúc dần thành hình trong đầu **anh** Tư .

Cường hỏi : “ thâm tâm anh có nghi ngờ vợ **anh** không ? “ .

POS Tags

- ▶ PUNCT: punctuation

Ví dụ:

- ▶ English:

Period: .

Comma: ,

Parentheses: ()

- ▶ Vietnamese:

giá nấu thuê 5 - 10 triệu đồng / nồi tùy ở thân sơ và cam kết chủ thợ .

- ▶ CONJ: A subordinating conjunction is a conjunction that links constructions by making one of them a constituent of the other

Ví dụ:

- ▶ English:

I believe **that** he will come

- ▶ Vietnamese:

đề nghị anh Đại liên hệ với chúng tôi : công ty Xây lắp & vật tư xây dựng , 93 Nguyễn công trú , q.1 , TP HCM .

POS Tags

- ▶ VERB: verb

Ví dụ:

Sau đó cảnh sát Pháp bắt tôi và dẫn tôi **đi** giam ở nhà giam Tours

Một hôm tôi biết **có** đoàn tàu chở vũ khí , xe tăng sang VN .
tôi **được** cử sang VN để giúp người VN chống phát xít Nhật .
chuyến bay sang châu Âu **phải** hoãn lại do thời tiết xấu .

- ▶ X: other: The tag X is used for words that for some reason cannot be assigned a real part-of-speech category. It should be used very restrictively.

Ví dụ:

Họ đã để **lại** hậu quả quá lớn bằng chất độc da cam
Nếu Tùng học **được** hai tuần nghiêm túc thì mẹ mua xe .
Tôi sinh **ra** trong gia đình nghèo .